14559ROUS01U

UNITED STATES PATENT APPLICATION FOR GRANT OF LETTERS PATENT

Kelvin Kar-Kin Au
Gamini Senarath
Yoon Chae Cheong
Wei Huang
INVENTORS

SCHEDULER WITH FAIRNESS CONTROL AND QUALITY OF SERVICE SUPPORT

Withrow & Terranova, P.L.L.C.
P.O. Box 1287
Cary, NC 27512
(919) 654-4520

Docket No. 7000-107

SCHEDULER WITH FAIRNESS CONTROL AND QUALITY OF SERVICE SUPPORT

Field of the Invention

[0001] The present invention relates to wireless communications, and in particular to scheduling data for transmission from a base station to one or more mobile terminals.

Background of the Invention

[0002] Wireless communication networks that allocate communication resources, such as time or frequency, require a scheduler to select data to be transmitted. When multiple users are vying for these resources, the scheduler must analyze the incoming data and determine the data having the highest priority for transmission. Priority has traditionally been based on maximizing overall system throughput or maintaining a certain Quality of Service (QoS) level to ensure data is transmitted in a timely fashion. When maximizing throughput, users having better channel conditions are favored over those with worse channel conditions. Thus, the users with the less favorable channel conditions are always given less priority unless time-sensitive data is discovered.

[0003] QoS analysis has been provided in existing systems but in a limited manner, for example, by processing only the section of the data that is ready for transmission or utilizing the buffer size. Normally, incoming data is buffered in queues corresponding to the user. If the QoS analysis only encompasses the section of data that is ready for transmission and not the entire queue, the remaining data in the queue does not affect priority. Thus, the scheduler cannot address urgent packets until they are ready for transmission, which results in QoS failures and inefficiencies. On the other hand, a large buffer size does not necessarily mean that all the packets in the buffer are urgent. These failures and inefficiencies typically further impact those users with less favorable channel conditions.

[0004] There have been some schemes to achieve a fixed level of fairness and maximize throughput. However, they lack: (a) the ability to control the fairness, including unfairness among users, (b) delay guarantee measures and (c) the ability to consider adequate information for data in the queue. Wireless network operators may want to control the amount of fairness in their systems. Given the inherent fixed level of fairness of the present schedulers, there is a need for a scheduler capable of maximizing the aggregate throughput, achieve the desired level of fairness among users and, at the same time, meet QoS requirements.

Summary of the Invention

[0005] The present invention provides a scheduler capable of maximizing aggregate throughput while achieving a controlled amount of fairness among users and meeting Quality of Service (QoS) requirements. The scheduler is configured to select the next unit of data to transmit from multiple queues associated with access terminals waiting to receive the data. For each access terminal, a weighting factor is calculated based on a temporal fading factor, a throughput fairness factor, and a delay QoS factor. The unit selected for transmission corresponds to the access terminal having the greatest overall weighting factor. The process repeats for each unit being transmitted.

[0006] The temporal fading factor is a measure of an access terminal's current channel conditions relative to the access terminal's average channel conditions. Although channel condition measurement may be derived using any number of techniques, one embodiment of the present invention monitors channel conditions based on carrier-to-interference (C/I) ratios provided by the access terminal. The term "current channel condition" can be the C/I ratio obtained from the access terminal or a compensated C/I ratio based on any compensation techniques, such as prediction. The throughput fairness factor is a function of throughput capability for each access terminal. In an effort to achieve a desired minimum degree of fairness for all users, the throughput fairness factor increases the priority of access terminals with lower throughput capability and decreases the priority of access terminals with higher throughput capability. The throughput fairness factor may be a function of actual throughput, channel conditions, or a combination thereof.

[0007] The delay QoS factor is a function of delay bound for a unit or series of units. The delay bound typically defines the time in which a unit or series of units must be delivered. The function is suitably optimized to account for the packets' delay bounds and the amount of data to transmit. In operation, the units in each queue are analyzed and given a weight inversely proportional to the delay bound. The weights for the units in a given queue are summed to arrive at the delay QoS factor. The delay QoS factor may be further adjusted based on the amount of data to transmit. Preferably, all of the units in each queue are weighted when calculating the delay QoS factor. To reduce processing, units having a delay bound greater than a defined limit may be assigned a nominal weight.

[0008] Those skilled in the art will appreciate the scope of the present invention and realize additional aspects thereof after reading the following detailed description of the preferred embodiments in association with the accompanying drawing figures.

Brief Description of the Drawing Figures

[0009] The accompanying drawing figures incorporated in and forming a part of this specification illustrate several aspects of the invention, and together with the description serve to explain the principles of the invention.

[0010] Figure 1 is a block representation of a wireless communication environment according to one embodiment of the present invention.

[0011] Figure 2 is a flow diagram according to one embodiment of the present invention.

[0012] Figures 3A and 3B are graphs illustrating exemplary current-to-mean channel conditions for two users, respectively.

[0013] Figures 4A and 4B are graphs illustrating both fair and unfair user throughput characteristics.

[0014] Figure 5 is a graph illustrating an exemplary weighting factor characteristic as a function of time.

Detailed Description of the Preferred Embodiments

[0015] The embodiments set forth below represent the necessary information to enable those skilled in the art to practice the invention and illustrate the best mode of practicing the invention. Upon reading the following description in light of the accompanying drawing figures, those skilled in the art will understand the concepts of the invention and will recognize applications of these concepts not particularly addressed herein. It should be understood that these concepts and applications fall within the scope of the disclosure and the accompanying claims.

[0016] Reference is now made to Figure 1. Wireless networks use access points, such as base stations 10, to facilitate communications with access terminals, such as mobile terminals 12, within a select coverage area, or cell. Respective groups of base stations 10 are supported by a communication network 14, which may include mobile switching centers, a public switched telephone network (PSTN), a packet-switched network, or a combination thereof. The communication network 14 is used to transport packets to and from the base station 10. The packets may be communicated in a direct packet-switched manner or on top of a circuit-switched platform. The manner in which the packets are communicated to the base station 10 is not critical to the invention.

[0017] During downlink communications from the base station 10 to select mobile terminals 12, the base station 10 must determine the manner and order in which to transmit the data received in the packets from the communication network 14 to the mobile terminals 12. Accordingly, the base station 10 will include a control system 16 having control plane 18 controlling the flow of data through a data plane 20. For communicating with the mobile terminals 12, the data plane 20 will process packets received from the communication network 14 via a network interface 22 under the control of the control plane 18. The packets are processed into units, which are delivered to radio frequency (RF) transceiver circuitry 24 for transmission. For the sake of clarity, the term “packet” refers to packetized data, which is received by the base station 10 from the communication network 14. The term “unit” refers to packetized data that is transmitted from the base station 10 to the mobile terminals 12. A unit may include all or any part of one or more packets. Although units may directly correspond to packets, units are preferably a given size wherein packets may vary in size from one packet to another. The units may include voice or traditional data.

[0018] The forward link from the base station 10 to the mobile terminal 12 will include one or more channels, which are divided into defined time slots. The RF transceiver circuitry 24 is configured to modulate a given unit as dictated by the control plane 18 and transmit the modulated unit via one or more antennas 26 during a single time slot. The RF transceiver circuitry 24 is preferably configured to implement different modulation and coding techniques and speeds based on channel conditions, the capabilities of the mobile terminals 12, or required transmission standards. Those skilled in the art will recognize the various possible modulation techniques and that multiple units may be transmitted in a given time slot

[0019] The control plane 18 includes a scheduler 28, which is configured to control delivery of units to the mobile terminals 12 based on channel conditions, required throughput fairness, and quality of service (QoS). During operation, packets for any number of mobile terminals 12 are received and stored in a buffer 30 associated with the data plane 20. The buffer 30 is segregated into multiple queues, each associated with a given mobile terminal 12. If the packets do not directly correspond to units, the incoming packets are processed into the desired units. The units are stored in the respective queues in the order in which they are received. Preferably, the queues use a first-in-first-out (FIFO) configuration.

[0020] With reference to the flow diagram of Figure 2, operation of the scheduler 28 is illustrated. On an ongoing basis, the units to transmit are placed in queues for the corresponding mobile terminals 12 (step 100). Further, the scheduler 28 will continuously monitor channel conditions and the throughput rates for each mobile terminal 12 (steps 102 and 104). A channel condition represents the quality of the transmission channel from the base station 10 to the mobile terminals 12. The throughput rates may be a function of actual data throughput, channel conditions, or a combination thereof.

[0021] Channel conditions may vary continuously and be determined using any number of techniques. For example, carrier to interference ratios (C/I), which represent a measure of signal power to interference power, may be fed back to the base station 10 from the mobile terminals 12. The scheduler 28 will preferably continuously track channel conditions and maintain an average channel condition over a select period of time as well as a current channel condition for each mobile terminal 12. Similarly, the scheduler 28 will preferably keep track of an average and current rate of data throughput for each of the mobile terminals 12.

[0022] The scheduler 28 continuously analyzes the units in each of the queues to determine the next unit to transmit. For each unit transmitted, the following processing occurs. The scheduler 28 will initially determine the current channel condition for each mobile terminal 12 being supported (step 106). From the current channel condition determination, a temporal fading factor is calculated for each mobile terminal (step 108). As noted, the C/I or like channel condition measurement may be used to determine the current channel condition. The temporal fading factor is preferably a measure of the current channel condition relative to an average channel condition for each mobile terminal 12.

[0023] The temporal fading factor increases for a mobile terminal 12 having the best current channel conditions relative to its own mean. Hence, each mobile terminal 12 will have a greater chance to receive units from the base station 10 than in a scenario where selections are biased toward mobile terminals 12 having the best current or average channel conditions. With reference to Figures 3A and 3B, the mobile terminal 12 of user 2 will have a greater temporal fading factor at time *t* than user 1, because the ratio between the current to the mean channel condition is higher for user 1 than user 2. Assigning mobile terminal 12 for user 2 a greater temporal fading factor helps to compensate for the historically lower channel conditions and the relatively high current channel condition with respect to the mobile terminal 12 of user 1. The temporal fading factor (T) may be implemented by:

$$T = \frac{\text{current C/I}}{\text{average C/I}}$$

or

$$T = \text{current C/I (dB)} - \text{average C/I (dB)}.$$

The average C/I is obtained over an averaging window, which is preferably approximately 20 fade duration. Fade duration is defined as $2f_d^{-1}$ where f_d is the maximum Doppler frequency. For instance, in the nomadic case where f_d = five (5) Hz, the averaging window is two (2) seconds.

[0024] Next, the scheduler 28 will calculate a throughput fairness factor for each mobile terminal 12 (step 110). The throughput fairness factor determines the amount of priority to give to mobile terminals 12 in locations supporting higher throughput versus those in locations supporting lower throughput. Although networks are typically configured to maximize throughput and to err toward maximizing throughput by prioritizing those mobile terminals 12 capable of communicating at higher rates, all users deserve at least a minimum degree of fairness, even at the expense of capacity. The throughput fairness factor is used to control the throughput distribution among the mobile terminals 12. With reference to Figures 4A and 4B, a more fair distribution is illustrated in Figure 4A while a seemingly unfair distribution is illustrated in Figure 4B. The scheduler 28 will generate a higher throughput fairness factor to allow those mobile terminals 12 with less favorable channel conditions to achieve a desired mean throughput.

[0025] The throughput fairness factor (F) may be implemented by $F_i = R_i^f$, where R_i is the average throughput capability of mobile terminal i (12) and f controls the desired fairness. Preferably, the average throughput capability is obtained from average channel conditions and not just on the throughput rate during transmission to a given mobile terminal 12. Accordingly, R_i is a function of channel conditions for each user i in one embodiment. Those skilled in the art will recognize that actual and average data throughput may be used to supplement determination of the fairness factor. In the present example, changing the parameter f provides various levels of fairness. For instance, when f is greater than zero, the throughput fairness factor F is larger for higher rate users and allows more unfairness. In contrast, when f is less than zero, the throughput fairness factor F is smaller for higher rate users and allows more fairness.

[0026] The scheduler 28 will next determine the requisite delivery time for the units in each of the queues associated with the mobile terminals 12 (step 112). Delay bounds relate to the time in which a unit (or packet) must be delivered to ensure a defined Quality of Service and may be used to determine expiry times. A delay bound is typically associated with a packet, which may be broken into one or more units. By analyzing all or a significant number of units in each queue, the scheduler 28 can make weighting decisions based on the urgency of delivery for packets deeper in the queue, even when the next unit to transmit in the queue does not have an impending delay bound.

[0027] After determining the expiry times for the queued units, the scheduler 28 will calculate a delay QoS factor (D) for each mobile terminal 12 (step 114). The delay QoS factor (D) for each mobile terminal 12 is preferably based on all of the units in the queue for the mobile terminal 12. In the example provided, each packet in each queue is given a weight corresponding to the inverse of the delivery time, which is the expiry time less the current time. A unit that is closer to its expiry time is given a larger weight than a unit further from its expiry time. To reduce computational load on the scheduler 28, units with a delivery time (expiry time – current time) greater than a given threshold or trigger, T_{TRIG}, may be assigned a normalized weight. Based on the weighting for each unit in the queue, a delay QoS factor (D) is calculated for the units at the head of each queue. Although various weighting criteria may be used, the following describes a weighting function for one embodiment of the present invention.

[0028] As illustrated in Figure 5, the present embodiment assigns a normalized weight for units having a remaining delivery time greater that T_{TRIG}. This delay QoS factor is calculated for each user at the beginning of each time slot by a weighted sum of all packets in the queue for all QoS levels. The weighted QoS factor for the i-th user is given by:

$$D_i(n) = \sum_{j\,=\,1\,-\,\text{all units in queue}} \frac{T_{TRIG} \text{ * Amount of data (unit or packet)}}{\text{(deliver time (j) - current time)}}$$

for (deliver time (j) – current time) < T_{TRIG}

[0029] The delay QoS factor (D) is a function of the amount of data and the requisite time to deliver as defined by the delay bounds. Therefore, the delay QoS factor (D) increases with increasing traffic and urgency of the units. Moreover, the delay QoS factor (D) can be large if there is a large amount of units far from the delay bound. Accordingly, the scheduler 28 can give priority to users with large amounts of data ahead of time.

[0030] Each unit in the queue is given a weight, which is inversely proportional to the remaining delivery time. A unit that is close to the expiry time has a larger weight than a unit that has plenty of time before it meets the delay bound. Only units that have a remaining delivery time greater than zero are counted. Further, the amount of data or size of a packet represented by one or more units will also impact the delay QoS factor (D). If the remaining delivery time for a unit is greater than T_{TRIG}, a normalized weighting is applied and D is simply the sum of all the packets. In essence, the parameter T_{TRIG} determines when the proportional weighting is applied to a given unit.

[0031] The scheduler 28 will then select the next unit to transmit based on the temporal fading factor (T), the throughput fairness factor (F), and the delay QoS factor (D) for the units at the head of each queue (step 116). One way of combining the three factors is to define a priority weighting index (P) based on the product of the three:

$$P = n\,(T_i\,)^x\,(F_i\,)^y\,(D_i)^z,$$

where T_i is the temporal fading factor, F_i is the throughput fairness factor, D_i is the delay QoS factor for the *ith* user, x, y, z are powers suitably chosen and n is a normalization constant. Accordingly, the product of the three factors is calculated for the units at the front of each queue. The unit with the greatest weighting factor is selected for delivery. Once selected, the unit at the front of the queue having the greatest weighting is modulated and transmitted during the next time slot by the RF transceiver circuitry 24 (step 118). The continuous process repeats for each available time slot.

[0032] Those skilled in the art will recognize improvements and modifications to the preferred embodiments of the present invention. All such improvements and modifications are considered within the scope of the concepts disclosed herein and the claims that follow.